DOCUMENT RESUME

ED 055 095 TM 000 811

AUTHOR Novick, Melvin R.; And Others

TITLE Applications of Bayesian Methods to the Prediction of Educational Performance.

INSTITUTION American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.

SPONS AGENCY National Inst. of Child Health and Human Development (NIH), Bethesda, Md.

PUB DATE Apr 71

NOTE 25p.

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS Academic Achievement; *Academic Performance; *Bayesian Statistics; Colleges; *College Students; Mathematical Applications; Multiple Regression Analysis; *Prediction; Predictive Ability (Testing); *Predictive Measurement; Predictor Variables; Sampling; Statistical Analysis

IDENTIFIERS *American College Testing Program

ABSTRACT

The feasibility and effectiveness of a Bayesian method for estimating regressions in m groups is studied by application of the method to data from the Basic Research Service of The American College Testing Program. Evidence supports the belief that in many testing applications the collateral information obtained from each subset of m-1 colleges will be useful for the estimation of the regression in the m-th college. Specifically, on cross-validation in a second sample, the Bayesian predictions had a smaller mean squared error in each of 22 colleges, the reduction averaging 9.7%, when compared with the least squares predictions when four predictor variables were used on a quarter sample in the 22 colleges where initial within-college sample sizes ranged from 26 to 184. Furthermore, even when based on the full sample within each college, the least squares predictions had an average cross-validated mean squared error only barely less than the Bayesian predictions based on the quarter sample. The most apparent benefit of the Bayesian method is that it permits regression to be done in subpopulations where sample sizes are small and where the regressions are different in the subpopulations. In the present study, a decrease of more than 10% in mean squared error was obtained using this approach. (Author)

ED055095

ACT RESEARCH REPORT

TM

No. 42

42

April 1971

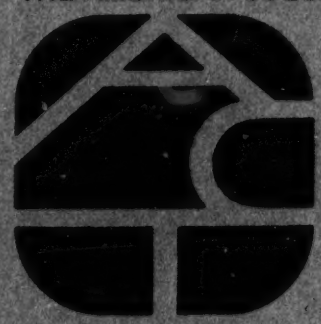
U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

APPLICATIONS OF BAYESIAN METHODS TO THE PREDICTION OF EDUCATIONAL PERFORMANCE

Melvin R. Novick
Paul H. Jackson
Dorothy T. Thayer
Nancy S. Cole

PUBLISHED BY THE RESEARCH AND DEVELOPMENT DIVISION

THE AMERICAN COLLEGE TESTING PROGRAM



P. O. BOX 168, IOWA CITY, IOWA 52240

TM 000 811

1

ABSTRACT

The feasibility and effectiveness of a Bayesian method, due to Lindley, for estimating regressions in m groups is studied by application of the method to data from the Basic Research Service of The American College Testing Program. Evidence is found to support the belief that in many testing applications the collateral information obtained from each subset of $m-1$ colleges will be useful for the estimation of the regression in the m-th college. Specifically, on cross-validation in a second sample, the Bayesian predictions had a smaller mean squared error in each of 22 colleges, the reduction averaging 9.7%, when compared with the least squares predictions when four predictor variables were used on a quarter sample in the 22 colleges where initial within-college sample sizes ranged from 26 to 184. Furthermore, even when based on the full sample within each college, the least squares predictions had an average cross-validated mean squared error only barely less than the Bayesian predictions based on the quarter sample. The most apparent benefit of the Bayesian method is that it permits regression to be done in subpopulations (e.g., male-female) where sample sizes are small and where the regressions are different in the subpopulations. In the present study, a decrease of more than 10% in mean squared error was obtained using this approach.

APPLICATIONS OF BAYESIAN METHODS TO THE PREDICTION OF EDUCATIONAL PERFORMANCE[1]

Melvin R. Novick
Paul H. Jackson
Dorothy T. Thayer[2]
Nancy S. Cole

Since 1961 The American College Testing Program has provided predictive research services to participating institutions. Applications of this service have been primarily in academic areas. A component of these services is the computation of a least squares regression equation for predicting grade point average (GPA) from a linear combination of the four scores, English, Mathematics, Social Studies, and Natural Sciences, on the American College Test and several high school grades. These regression equations are computed on first semester GPA and then used on different students for predictive purposes for the next year. Predictive information is furnished to the college and to each applicant in order to foster a more rational, mutual evaluation of the benefits a particular student might gain from a particular college. The selection of students by a college and the selection of a college by students depends on many factors other than predicted GPA, but this prediction is one objective and very useful piece of information.

A requirement for participation in the predictive research services has been the availability of a minimum sample of 100 students in a single year. This arbitrary number was set as a result of experience suggesting that smaller samples typically provide an unsatisfactory amount of sampling fluctuation. Unfortunately, many smaller colleges are not able to supply the requisite sample. Also, there is much interest in doing separate regression analyses for subgroups of different kinds of students and different kinds of programs. Such analyses necessarily involve smaller sample sizes. The present extension of national testing programs into the areas of vocational and technical education intensifies this problem and provides a challenge that can be met only by a radically new and improved prediction technology.

---

[1] An invited paper presented for discussion at the American Statistical Association meetings, Detroit, Michigan, December 27, 1970. The authors are grateful to Mr. David Christ for preparing the computer programs used in the cross-validation analyses reported here, and to Dr. E. James Maxey for providing the data from the ACT Basic Research Services file. Supported in part under Grant 1 P01 HD01762 from the National Institute of Child Health and Human Development. Reproduction, translation, use, or disposal by or for purposes of the U.S. Government is permitted. The substance of this report is also contained in Research Bulletin 71-18 of the Educational Testing Service. The final published report will contain further theoretical details.

[2] Dorothy T. Thayer is an assistant statistician with the Educational Testing Service.

In vocational-technical programs, available criterion groups within specific programs will regularly be of very small size, much smaller than normally thought necessary for accurate prediction. One can get larger samples by pooling related programs, but this simple pooling is unlikely to be very satisfactory. However, a modified kind of pooling of students from like named programs from different institutions is possible and, indeed, may be useful. We recognize that regression weights for similar named programs should be similar across institutions, but we also know from experience that they will be different enough that direct pooling will be less than entirely satisfactory. What is really needed is a technique that takes account both of the similarity of regressions across institutions and also the uniqueness of the individual programs.

A simple paradigm for a powerful approach to this problem has classic status in mental testing theory. Suppose we administer just two items randomly selected from the ACT Mathematics subtest and from a student's two incorrect responses we project a scaled ACT Mathematics score of 10. Further suppose we know that the student comes from a high school in which the average score obtained by students is 25. The two pieces of information seem to conflict. If we had only the test results we would certainly want to use the projected value of 10. However, if we had no information on the student but only knew that he came from a school where the mean score was 25, we would probably pick that value as our best guess of his true ability, his true score.

In the present intermediate situation it would seem to make sense to use both pieces of information giving appropriately different weights to each. Kelley (1927) provided a formal solution to this problem some 40 years ago. His solution was to weight the observed score for the person and the average value over persons, respectively, by the reliability of the test and one minus the reliability. Symbolically this is expressed

$$\text{Estimated true score} = rX + (1 - r)\overline{X}$$

where X is the person's observed score, $\overline{X}$ is the mean of all observed scores, and r is the reliability of the test defined as the ratio of the variance of the true scores to the variance of the observed scores in the given population. The reliability of a pair of mathematics items is, perhaps, .30 and so, by this formula, we would have:

$$\text{Estimated true score} = (.30)(10) + (.70)(25) = 20.50.$$

Had the observed score been based on four items, the reliability would have been about .45 and

$$\text{Estimated true score} = (.45)(10) + (.55)(25) = 18.25.$$

Had the observed score been based on the full subtest the reliability would be perhaps .90, and

$$\text{Estimated true score} = (.90)(10) + (.10)(25) = 11.25.$$

This procedure seems to make a great deal of sense. The *collateral* information contained in the scores obtained by the other students from the same school is information that should be ignored only if the reliability of the test is very high. It is not surprising to learn that Kelley showed that unless the reliability is very high the above procedure provides an estimate with substantially lower standard error than the estimate based solely on the observed score.

The necessary breakthrough in prediction technology came when Lindley (1970) showed that the logic of the Kelley method could be used to improve predictions in individual groups by using information both from that group and from other similar groups. While the actual mechanics of the Lindley method are complicated, they effectively involve the same kind of averaging as in the estimation of true score. Suppose there are m colleges and p predictor variables; then for each of the predictors and for each college a Bayesian regression weight is computed as a weighted average of the usual least squares regression weight for that variable and that college and the average over colleges of the weights for that variable. For example, suppose we have 10 colleges with similar programs. In one college, we compute the following least squares regression:

Least Squares Regression Weights (LSRW) for College 1

Variables	1	2	3	4
Weights	.02	.01	.02	−.01

We make a similar computation with each college and for each variable separately obtain the average over colleges of these weights.

Average of LSRW over 10 Colleges

Variables	1	2	3	4
Weights	.03	.03	.02	.01

The Bayesian regression weights for College 1 would then be computed as the following weighted averages:

$$(.02)(W_1) + (.03)(1 - W_1)$$
$$(.01)(W_2) + (.03)(1 - W_2)$$
$$(.02)(W_3) + (.02)(1 - W_3)$$
$$(-.01)(W_4) + (.01)(1 - W_4)$$

The weights W_1, W_2, W_3, and W_4 depend (among other things) upon the relative amount of information we have from College 1. If we have a lot of information on this college, the weights (W_i) will be large and the Bayesian estimates will be similar to the least squares regression weights for College 1. If, however, we have only a little information about College 1, the weights (W_i) will be small and the Bayesian regression weights will be more like the average for all 10 colleges. Suppose the weights were

$$(W_1, W_2, W_3, W_4) = (.3, .4, .4, .4)$$

Then the "Kelley" regression weights for College 1 would be .027, .022, .020, .002. The actual Bayesian procedure is much more complicated, but conceptually it differs little from the simple technique described above.

Now we know that the least squares regression weights minimize the error in the present sample of students from College 1. Use of the Bayesian weights in the present sample would result in larger errors. But what will happen with next year's students? Which type of regression weights actually predicts better for next year's students? In this paper we report on an experiment that demonstrates the great value of the Bayesian regression weights. We present, finally, a technical section giving further details on Lindley's method of deriving these weights.

The Cross-Validation Experiment

In theory the Bayesian regression estimates should provide "better cross-validation" than the usual least squares estimates. This would mean that predictions using the Bayesian estimates would be more accurate, on the average, than those using the least squares estimates. Typically comparative studies are done on computer generated data. Such studies are useful, but never conclusive. These studies can tell us that the theory does or does not work when the assumptions of the model are satisfied or violated in specific ways. However, only a study with real data can pin down precisely how a technique will work on further data sets.

In the fields of education and psychology, the process of seeing just how a prediction equation works in a second sample is called *cross-validation.* The variable being predicted is called the *criterion,* and the independent variables are called the *predictors.* The typical finding is that when the regression equation from one sample is used in a second sample the correlation between predicted values and the observed value of the criterion is lower, on the average, than it was in the first sample. This is a result of the fact that the first sample regression equation imprecisely estimates the true relationship because it fits idiosyncrasies found in the first but not the second sample. With samples that are small relative to the number of predictors used, very high multiple correlations can even cross-validate to zero. While one can devise formulas to predict this shrinkage, it seems far more useful to reduce the shrinkage by discounting, to the extent possible, idiosyncrasies in the first sample. This is precisely the function of the Bayesian method.

For the present study, we decided to work with a group of traditional junior colleges providing academic transfer curricula. A sample of 22 such colleges was drawn from those participating in the Basic Research Service during both 1968 and 1969. The sample was selected without reference to the present data but with careful attention to the curricula in the colleges, using information from the standard reference *American Junior Colleges* (Gleazer, 1967). We consider this to be a very homogeneous group of colleges since colleges that appeared at all different from the other members of the group were eliminated. On the information available to us, we consider these and similar colleges to be exchangeable, as regards the important considerations of this study. A fundamental statistical theorem, the de Finetti-Hewitt-Savage theorem, implies that we may, therefore, treat these as if they were a random sample from some population of colleges. This, together with the usual tractability considerations, implies precisely the variance components model, Model II, to be considered here.

There are several common measures for comparing predicted GPA with observed GPA. The most common is mean squared error (MSE), the average of the squares of the differences between the person's predicted GPA and his actual GPA, within each college. If, however, the principal interest is in ranking students within each college, the correlation (COR) between the observed and predicted grades is a more suitable basis for comparison. Using COR, differences in the mean observed and mean predicted grades are ignored. On the other hand, another possible concern is with the average absolute magnitude of the error of prediction (AE) for each college. Finally, for some types of college decisions, a zero-one loss function (ZOL) is best. Differences in observed and predicted grades within a certain range may be acceptable (zero loss) while those larger are considered to be errors (unit loss). In this case, in order to keep the measures comparable from one college to another, the acceptable interval was defined as one-half a standard deviation of observed grades within each college. These four measures of goodness of prediction were computed for each of the 22 junior colleges. An overall index for each measure was obtained by averaging it over the 22 colleges.

The above computations were made using the least squares and the Bayesian linear prediction functions. The number of observations within the colleges ranged from 105 to 739. In addition to making comparisons using all of the 1968 data, there was a need to make similar comparisons with smaller within-college sample sizes. For this purpose, a 25% random sample within each college was drawn and the entire study redone. The number of cases within the various colleges ranged from 26 to 184 in the 25% sample. Our overall expectation was that the Bayesian method would provide a modest average improvement with the full sample and a much more substantial improvement with the quarter sample.

Following this, the full 1968 sample in each college was divided into male and female subpopulations, and the least squares and Bayesian estimates were computed in each college sex subgroup. A comparative

cross-validation analysis was then performed in male-female groups in the 1969 sample. The expectation was that the greatest benefit would be found in this application.

Results

In Table 1 we give the results of the least squares regression analyses within each of the 22 colleges for the 1968 data. The symbol $\hat{\beta}_0$ refers to the sample intercept and $\hat{\beta}_1$, ..., $\hat{\beta}_4$ are the four sample least squares regression coefficients corresponding to the four ACT test scores, English, Mathematics, Social Studies, and Natural Sciences, respectively. The sample multiple correlations with GPA, which range between .4105 and .6882, are given in the column labeled R. The column labeled $\hat{\phi}$ gives the estimated residual variances. Although there are clearly differences among the institutions, our thought that they are very similar is surely confirmed by these sets of estimates.

TABLE 1

Regression Coefficients and Residual Variances for ACT Test Predictions of GPA in 22 Colleges (100% Sample, 1968 Data)

COLL	N	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R	$\hat{\phi}$
1	171	0.798	0.023	0.006	0.037	0.011	.4612	0.57493
2	204	1.172	0.042	0.011	0.006	−0.005	.4765	0.29663
3	223	0.317	0.043	0.021	0.030	0.018	.6234	0.37718
4	307	0.925	0.009	0.036	0.003	0.016	.4507	0.47905
5	461	0.775	0.033	0.019	0.011	0.011	.4561	0.38946
6	175	0.104	0.029	0.036	0.020	0.007	.4865	0.58086
7	105	0.287	0.069	0.025	0.007	0.001	.5237	0.58290
8	118	0.362	0.005	0.037	0.042	0.004	.5402	0.52335
9	113	0.401	0.042	0.033	0.042	−0.023	.5429	0.49640
10	128	0.045	0.068	0.038	0.012	0.011	.6882	0.44949
11	165	1.087	0.060	0.006	0.022	−0.005	.5502	0.42804
12	132	1.205	0.048	0.005	0.012	0.007	.5478	0.29100
13	174	0.916	0.024	0.025	0.002	0.017	.4905	0.36637
14	334	0.122	0.070	0.019	0.017	0.001	.5095	0.83703
15	167	0.215	0.070	0.031	0.026	−0.020	.5775	0.40428
16	327	0.385	0.043	0.034	0.015	−0.006	.4745	0.44930
17	739	0.864	0.035	0.026	0.014	−0.013	.4317	0.39968
18	235	1.193	0.043	0.006	0.025	−0.009	.4105	0.44821
19	117	1.056	0.065	0.006	0.024	−0.009	.4756	0.38558
20	209	0.892	0.053	0.007	0.010	0.012	.4438	0.48622
21	394	−0.277	0.078	0.034	0.011	−0.008	.5207	0.62199
22	410	0.075	0.030	0.046	0.011	0.013	.5591	0.46034

The multiple correlations found in the 1968 sample are typical of the results to be found when the ACT test is used in colleges having a homogeneous first semester program. Obviously, when students take widely varying kinds of courses during their first semester no single regression equation will be suitable, but rather, specific regression equations must be computed for each program. If this is not done, multiple correlations using fixed combinations of predictors can be very low.

The striking feature of this table is the presence of negative regression weights for the Natural Science variable ($\hat{\beta}_4$) in 9 of the 22 colleges. We must ask ourselves if these are accurate estimates or if the correct values are all zero or slightly positive, with the negative sample quantities due solely to sampling variation. The data are certainly consistent with this hypothesis on first glance. We look to the Bayesian analysis to clarify this point.

In Table 2 we give the least squares estimates together with the Bayesian estimates for each college. The general effect has been to moderate extreme values throughout. The regression weight for Natural Science (β_4) is estimated to be near zero (actually just slightly positive) for all colleges. No meaningful differences in the regression coefficients, across colleges, were detected for the Social Studies (β_3) or Natural Sciences (β_4) variables, but substantial differences were found for the English (β_1) and Mathematics (β_2) variables. A negative intercept (β_0)is still found for College 21, but this is small enough that it does not cause us undue discomfort.

TABLE 2

Comparison of Least Squares and Bayesian Prediction Functions
(100% Sample, 1968 Data)

COLL	N		β_0	β_1	β_2	β_3	β_4
1	171	LSQ	0.7981	0.0231	0.0062	0.0374	0.0107
		BAY	0.8374	0.0406	0.0195	0.0169	0.0017
2	204	LSQ	1.1724	0.0423	0.0114	0.0060	−0.0046
		BAY	0.9737	0.0295	0.0175	0.0169	0.0017
3	223	LSQ	0.3169	0.0430	0.0208	0.0300	0.0181
		BAY	0.6152	0.0562	0.0223	0.0169	0.0017
4	307	LSQ	0.9254	0.0094	0.0363	0.0032	0.0164
		BAY	0.8276	0.0293	0.0235	0.0169	0.0017
5	461	LSQ	0.7753	0.0329	0.0192	0.0112	0.0108
		BAY	0.7605	0.0349	0.0219	0.0169	0.0017
6	175	LSQ	0.1041	0.0293	0.0359	0.0200	0.0070
		BAY	0.1599	0.0383	0.0334	0.0169	0.0017
7	105	LSQ	0.2864	0.0686	0.0250	0.0075	0.0006
		BAY	0.4192	0.0494	0.0262	0.0169	0.0017

continued

TABLE 2 (continued)

COLL	N		β_0	β_1	β_2	β_3	β_4
8	118	LSQ	0.3619	0.0046	0.0367	0.0424	0.0042
		BAY	0.3742	0.0420	0.0291	0.0169	0.0017
9	113	LSQ	0.4011	0.0416	0.0332	0.0417	−0.0231
		BAY	0.4431	0.0467	0.0265	0.0169	0.0017
10	128	LSQ	0.0448	0.0681	0.0385	0.0123	0.0101
		BAY	0.4244	0.0631	0.0251	0.0169	0.0017
11	165	LSQ	1.0865	0.0596	0.0063	0.0221	−0.0051
		BAY	1.0643	0.0510	0.0131	0.0169	0.0017
12	132	LSQ	1.2053	0.0475	0.0045	0.0118	0.0068
		BAY	1.1415	0.0433	0.0128	0.0169	0.0017
13	174	LSQ	0.9157	0.0241	0.0253	0.0016	0.0174
		BAY	0.8352	0.0342	0.0212	0.0169	0.0017
14	334	LSQ	0.1218	0.0699	0.0187	0.0175	0.0014
		BAY	0.2375	0.0539	0.0285	0.0169	0.0017
15	167	LSQ	0.2148	0.0701	0.0309	0.0259	−0.0196
		BAY	0.3451	0.0519	0.0273	0.0169	0.0017
16	327	LSQ	0.3847	0.0434	0.0339	0.0153	−0.0062
		BAY	0.3299	0.0405	0.0296	0.0169	0.0017
17	739	LSQ	0.8642	0.0349	0.0264	0.0141	−0.0133
		BAY	0.7245	0.0289	0.0229	0.0169	0.0017
18	235	LSQ	1.1934	0.0425	0.0064	0.0251	−0.0091
		BAY	1.0253	0.0412	0.0151	0.0169	0.0017
19	117	LSQ	1.0554	0.0645	0.0065	0.0237	−0.0093
		BAY	1.0919	0.0517	0.0125	0.0169	0.0017
20	209	LSQ	0.8921	0.0528	0.0074	0.0100	0.0120
		BAY	0.8857	0.0468	0.0173	0.0169	0.0017
21	394	LSQ	−0.2766	0.0778	0.0338	0.0110	−0.0082
		BAY	−0.1400	0.0524	0.0358	0.0169	0.0017
22	410	LSQ	0.0751	0.0301	0.0461	0.0111	0.0128
		BAY	0.1698	0.0428	0.0344	0.0169	0.0017

The cross-validations using the 100% sample are given in Table 3. The advantage of the Bayesian method is very modest, being approximately 1% more or less, on the average, depending upon the error function considered. Thus, unless these results are contradicted in future samples or unless some refinements are found in the method, we must conclude that the Bayesian method has little to offer in the way of increasing predictive efficiency with sample sizes as large as those being considered.

TABLE 3

Comparisons of Classical and Bayesian Predictions of 1969 Data Using 1968 Data from 100% Sample

COLL	N		MSE	AE	ZOL	COR
1	171	LSQ	0.6652	0.6614	0.6145	0.3894
		BAY	0.6675	0.6721	0.6145	0.3827
2	204	LSQ	0.4469	0.5094	0.4854	0.5345
		BAY	0.4325	0.5034	0.5146	0.5417
3	223	LSQ	0.4322	0.5020	0.5400	0.4508
		BAY	0.4127	0.4881	0.5200	0.4706
4	307	LSQ	0.5559	0.5806	0.5683	0.3282
		BAY	0.5312	0.5662	0.5535	0.3805
5	461	LSQ	0.4887	0.5454	0.5373	0.4524
		BAY	0.4843	0.5429	0.5448	0.4598
6	175	LSQ	0.8589	0.7530	0.6034	0.5065
		BAY	0.8427	0.7448	0.6034	0.5229
7	105	LSQ	0.4486	0.5345	0.5652	0.4533
		BAY	0.4525	0.5373	0.5826	0.4397
8	118	LSQ	0.3653	0.4595	0.4821	0.5991
		BAY	0.3391	0.4438	0.4375	0.6366
9	113	LSQ	0.5761	0.6231	0.6176	0.4453
		BAY	0.5933	0.6187	0.6078	0.4047
10	128	LSQ	0.5495	0.5968	0.6525	0.3677
		BAY	0.4950	0.5723	0.6186	0.3757
11	165	LSQ	0.5767	0.5427	0.4437	0.5760
		BAY	0.5764	0.5447	0.4375	0.5801

continued

TABLE 3 (continued)

COLL	N		MSE	AE	ZOL	COR
12	132	LSQ	0.3883	0.4907	0.5463	0.5213
		BAY	0.3853	0.4959	0.5556	0.5278
13	174	LSQ	0.5967	0.6082	0.5570	0.4536
		BAY	0.5744	0.5958	0.5380	0.4881
14	334	LSQ	0.7164	0.6926	0.6446	0.4604
		BAY	0.7264	0.6986	0.6506	0.4494
15	167	LSQ	0.4876	0.5538	0.5928	0.3607
		BAY	0.4713	0.5418	0.5689	0.3548
16	327	LSQ	0.5836	0.5861	0.5246	0.4961
		BAY	0.5786	0.5826	0.5246	0.5031
17	739	LSQ	0.4632	0.5408	0.6222	0.4013
		BAY	0.4553	0.5373	0.6159	0.4016
18	235	LSQ	0.8013	0.7354	0.6061	0.5476
		BAY	0.7907	0.7282	0.6104	0.5520
19	117	LSQ	0.7549	0.6531	0.4818	0.5055
		BAY	0.7501	0.6411	0.4455	0.5221
20	209	LSQ	0.4301	0.5088	0.5376	0.5699
		BAY	0.4243	0.5063	0.5434	0.5740
21	394	LSQ	0.7252	0.6740	0.5680	0.4816
		BAY	0.7223	0.6737	0.5561	0.4800
22	410	LSQ	0.4008	0.5144	0.5594	0.4948
		BAY	0.3979	0.5137	0.5718	0.4961
AVE ERROR		LSQ	0.5596	0.5848	0.5614	
		BAY	0.5502	0.5795	0.5552	

9

We then undertook to study what would happen if the sample requirement were indeed reduced to a minimum sample size of 25. A random 25% sample within each college was used and the study redone. Sample sizes now ranged from 26 to 184. In Table 4, which is analogous to Table I, we give the least squares estimates for the 25% sample from each college. Notice that with these greatly reduced sample sizes, there are far more negative regression weights. They are even to be found for the first two predictors despite the fact that the 100% sample has strongly suggested that these weights should be positive. Furthermore, two of these negative weights are of substantial magnitude, specifically the Mathematics variable ($\hat{\beta}_2$) for College 1 and the English variable ($\hat{\beta}_1$) for College 8.

TABLE 4

Regression Coefficients and Residual Variances for ACT Test Predictions of GPA in 22 Colleges (25% Sample, 1968 Data)

COLL	N	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R	$\hat{\phi}$
1	42	0.841	0.023	−0.037	0.037	0.034	.4834	0.59344
2	51	1.533	−0.006	0.028	0.059	−0.053	.5504	0.27666
3	55	−0.348	0.052	0.036	0.010	0.038	.6138	0.43976
4	76	0.964	−0.010	0.055	0.004	0.015	.4749	0.40149
5	115	0.379	0.047	0.017	−0.001	0.031	.5060	0.39646
6	43	0.717	0.005	0.017	0.026	0.009	.3026	0.68759
7	26	−0.369	0.060	0.004	0.033	0.028	.6624	0.33275
8	29	1.182	−0.055	−0.002	0.121	−0.016	.6625	0.26714
9	28	−1.095	0.038	0.000	0.051	0.079	.8288	0.27665
10	32	0.670	−0.009	0.047	0.028	0.032	.7735	0.30405
11	41	1.051	0.064	0.000	0.012	0.012	.5369	0.45857
12	33	0.856	0.041	0.017	0.015	0.025	.7404	0.16812
13	43	0.797	−0.018	0.039	0.014	0.027	.4731	0.35890
14	83	0.128	0.034	0.041	0.035	−0.002	.5390	0.75269
15	41	0.571	0.023	0.017	0.022	0.018	.5230	0.33592
16	81	−0.469	0.094	0.036	0.005	0.003	.6293	0.46772
17	184	0.815	0.040	0.028	−0.004	0.000	.4322	0.39847
18	58	1.362	0.042	−0.003	0.046	−0.032	.4392	0.41840
19	29	0.812	0.045	0.007	0.038	0.007	.5453	0.39524
20	52	1.193	0.042	−0.002	0.006	0.021	.5125	0.38224
21	98	−0.409	0.052	0.035	0.028	0.005	.5260	0.58589
22	102	0.286	0.030	0.028	0.030	0.001	.4693	0.47440

The results of the Bayesian regression analyses are given in Table 5. Again we find that the negative regression weights have vanished. No differences across groups are found in the fourth regression weight and only very small differences in the third weight. The Bayesian estimates of the regression weights from the 25% sample are very similar to those from the 100% sample.

TABLE 5

Comparison of Least Squares and Bayesian Prediction Functions
(25% Sample, 1968 Data)

COLL	N		β_0	β_1	β_2	β_3	β_4
1	42	LSQ	0.8405	0.0227	−0.0370	0.0373	0.0336
		BAY	0.4923	0.0327	0.0229	0.0192	0.0111
2	51	LSQ	1.5331	−0.0062	0.0281	0.0591	−0.0525
		BAY	0.6596	0.0285	0.0224	0.0171	0.0111
3	55	LSQ	−0.3481	0.0519	0.0364	0.0101	0.0384
		BAY	0.5319	0.0364	0.0235	0.0192	0.0111
4	76	LSQ	0.9635	−0.0105	0.0546	0.0042	0.0155
		BAY	0.7754	0.0267	0.0231	0.0160	0.0111
5	115	LSQ	0.3786	0.0467	0.0165	−0.0011	0.0313
		BAY	0.5377	0.0338	0.0230	0.0178	0.0111
6	43	LSQ	0.7173	0.0051	0.0166	0.0261	0.0088
		BAY	0.3034	0.0307	0.0239	0.0183	0.0111
7	26	LSQ	−0.3690	0.0596	0.0044	0.0331	0.0276
		BAY	0.4234	0.0351	0.0234	0.0192	0.0111
8	29	LSQ	1.1819	−0.0545	−0.0020	0.1214	−0.0158
		BAY	0.6297	0.0312	0.0228	0.0185	0.0111
9	28	LSQ	−1.0948	0.0383	0.0000	0.0506	0.0790
		BAY	0.3404	0.0376	0.0238	0.0208	0.0111
10	32	LSQ	0.6702	−0.0094	0.0472	0.0276	0.0317
		BAY	0.8390	0.0330	0.0226	0.0181	0.0111
11	41	LSQ	1.0505	0.0640	−0.0003	0.0123	0.0122
		BAY	1.0468	0.0335	0.0213	0.0175	0.0111
12	33	LSQ	0.8564	0.0406	0.0170	0.0154	0.0248
		BAY	1.0310	0.0332	0.0216	0.0175	0.0111
13	43	LSQ	0.7968	−0.0177	0.0387	0.0144	0.0273
		BAY	0.5557	0.0301	0.0233	0.0175	0.0111

continued

TABLE 5 (continued)

COLL	N		β_0	β_1	β_2	β_3	β_4
14	83	LSQ	0.1281	0.0345	0.0409	0.0355	−0.0025
		BAY	0.4068	0.0369	0.0239	0.0204	0.0111
15	41	LSQ	0.5713	0.0232	0.0167	0.0218	0.0181
		BAY	0.4985	0.0331	0.0232	0.0186	0.0111
16	81	LSQ	−0.4690	0.0940	0.0362	0.0046	0.0026
		BAY	0.2706	0.0399	0.0243	0.0201	0.0111
17	184	LSQ	0.8150	0.0396	0.0282	−0.0041	−0.0004
		BAY	0.5875	0.0291	0.0230	0.0142	0.0111
18	58	LSQ	1.3618	0.0424	−0.0027	0.0459	--0.0322
		BAY	0.8372	0.0322	0.0215	0.0176	0.0111
19	29	LSQ	0.8125	0.0450	0.0066	0.0375	0.0073
		BAY	0.9686	0.0345	0.0217	0.0183	0.0111
20	52	LSQ	1.1932	0.0421	−0.0017	0.0062	0.0213
		BAY	0.8965	0.0316	0.0217	0.0168	0.0111
21	98	LSQ	−0.4091	0.0520	0.0353	0.0280	0.0050
		BAY	0.0946	0.0376	0.0243	0.0208	0.0111
22	102	LSQ	0.2856	0.0297	0.0278	0.0296	0.0009
		BAY	0.2967	0.0335	0.0240	0.0195	0.0111

Again, both the Bayesian and least squares equations were applied to the 1969 sample and the predictions compared. Some of the key cross-validation results of the study are found in Table 6. When the reduced sample was used to construct the prediction equations, the reduction in mean squared error (MSE) comparing the Bayesian method to the least squares method in the following year's students is about 9.7%, with smaller reductions in absolute (AE) and zero-one error (ZOL). Furthermore, some improvement in mean squared error was found in each of the colleges. The increase in the cross-validation correlations (COR) was about .04 on the average. We judge these to be very significant improvements indeed. The most impressive result, however, is in the comparison of the average mean squared errors for the Bayesian method with the 25% sample and the classical method with the 100% sample. A difference of less than one-quarter of 1% was found. Apparently, with these data, a 75% savings in sample size was possible by adopting the Bayesian method.

An interesting supplementary analysis was performed in which the prediction functions from the 1968 25% sample were cross-validated on the remaining (75%) 1968 sample. Average mean squared errors on this within-year cross-validation were: for least squares .5518 and for Bayes .5037 as compared to .6208 and .5603 when the cross-validation was done on the 1969 data. The gain using Bayes was 8.7% as compared with

TABLE 6

Comparisons of Classical and Bayesian Predictions of 1969 Data Using 1968 Data from 25% Sample

COLL	N		MSE	AE	ZOL	COR
1	42	LSQ	0.7717	0.7254	0.6704	0.3511
		BAY	0.7477	0.7189	0.6536	0.3643
2	51	LSQ	0.5579	0.5934	0.5848	0.3607
		BAY	0.4398	0.5228	0.5439	0.5383
3	55	LSQ	0.5051	0.5404	0.5360	0.4661
		BAY	0.4574	0.5259	0.5520	0.4704
4	76	LSQ	0.5985	0.6062	0.6089	0.2736
		BAY	0.5307	0.5665	0.5535	0.3736
5	115	LSQ	0.5098	0.5555	0.5653	0.4323
		BAY	0.4946	0.5480	0.5485	0.4522
6	43	LSQ	0.9511	0.8014	0.5866	0.4454
		BAY	0.8594	0.7476	0.5866	0.5123
7	26	LSQ	0.4809	0.5473	0.5739	0.4598
		BAY	0.4676	0.5478	0.6000	0.4259
8	29	LSQ	0.4819	0.5449	0.5179	0.4525
		BAY	0.3397	0.4379	0.4375	0.6409
9	28	LSQ	0.9632	0.7619	0.6275	0.3398
		BAY	0.6238	0.6387	0.6078	0.3992
10	32	LSQ	0.5804	0.6236	0.6780	0.2905
		BAY	0.4604	0.5613	0.6356	0.3592
11	41	LSQ	0.5918	0.5393	0.4375	0.5608
		BAY	0.5907	0.5470	0.4375	0.5579
12	33	LSQ	0.4152	0.4994	0.5463	0.4851
		BAY	0.3983	0.4996	0.5370	0.4973
13	43	LSQ	0.7184	0.6701	0.5759	0.3187
		BAY	0.6075	0.6178	0.5759	0.4705

continued

TABLE 6 (continued)

COLL	N		MSE	AE	ZOL	COR
14	83	LSQ	0.7495	0.7078	0.6596	0.4257
		BAY	0.7153	0.6883	0.6054	0.4320
15	41	LSQ	0.4881	0.5551	0.6048	0.3225
		BAY	0.4793	0.5490	0.6108	0.3351
16	81	LSQ	0.6188	0.5927	0.5014	0.4948
		BAY	0.5882	0.5840	0.5275	0.5048
17	184	LSQ	0.4880	0.5493	0.6127	0.3681
		BAY	0.4668	0.5421	0.6095	0.3898
18	58	LSQ	0.8239	0.7402	0.6147	0.4552
		BAY	0.7734	0.7212	0.6234	0.5174
19	29	LSQ	0.7442	0.6306	0.4727	0.5139
		BAY	0.7243	0.6223	0.4727	0.5167
20	52	LSQ	0.4849	0.5490	0.6012	0.5477
		BAY	0.4398	0.5208	0.5665	0.5595
21	98	LSQ	0.7325	0.6716	0.5609	0.4792
		BAY	0.7237	0.6805	0.5561	0.4757
22	102	LSQ	0.4020	0.5153	0.5817	0.4871
		BAY	0.3982	0.5162	0.5693	0.4965
AVE ERROR		LSQ	0.6208	0.6146	0.5781	
		BAY	0.5603	0.5866	0.5641	

9.7% when the cross-validation was done with the 1969 data. The suggestion here is that the Bayes procedure, to some extent, smoothes out year to year sample variations. These variations are an important cause of the shrinkage found in the multiple correlation when weights from one year are used in a second year. Thus, it is clear that only cross-validations on data from a subsequent year include all the important types of variation found in actual practice.

The effect of doing the predictions on the 1969 data separately for males and females from the corresponding 1968 data was to substantially reduce the average mean squared error. The relevant figures are given in Table 7.

It is clear from Table 7 that a worthwhile reduction in mean squared error can be obtained when Bayesian weights are used and prediction is done separately for males and females. Some reduction (3.6%) is found, on

TABLE 7

Comparison of Cross-Validated Average Mean Squared Errors Using Least Squares and Bayesian Predictions When the GPA Predictions Are Done Separately for Males and Females

ORIGIN OF WEIGHTS	GROUP PREDICTED FOR	LSQ	BAY
100% Sample Combining Sexes (1968)	1969 Combined Sexes	.5596	.5502
100% Sample Combining Sexes (1968)	1969 Males	.5641	.5539
100% Sample Combining Sexes (1968)	1969 Females	.5505	.5400
25% Sample Combining Sexes (1968)	1969 Combined Sexes	.6208	.5603
Males only from 100% Sample (1968)	1969 Males	.5609	.5418
Females only from 100% Sample (1968)	1969 Females	.5004	.4632

the average, using classical weights in the divided group, but this average reduction is much less than with the Bayesian weights (10%). The relative efficacy of the Bayesian method will be even greater when predictions are done in the subgroups of even smaller sample sizes. With present sample sizes, it appears that we can get a 10% reduction in mean squared error with predictions in the divided group *if* we use the Bayesian method.

At present, a minimum sample size of 100 is required for any regression function reported by the Basic Research Service, and thus many colleges are unable to have separate weights for males and females. With the new Bayesian methodology, it is clear that, for the kinds of colleges being studied here, requirements can easily be cut in half while providing better predictions than are now being provided. There seems little reason to doubt that this conclusion will hold for other testing programs.

We argue that a major benefit of the Bayesian method is that it permits working with smaller samples and that this, in turn, permits working in separate subpopulations. Not only does this increase overall efficiency of prediction but it provides a fairer prediction system. Consider the Bayesian prediction weights for College 17: for simplicity, suppose we set $x_1 = x_2 = x_3 = x_4$, then the prediction functions for males, females, and the overall population are as follows:

			x = 10	20	30
males	♂	.693 + .063x	1.32	1.95	2.58
females	♀	.670 + .088x	1.55	2.43	3.31
overall	□	.725 + .071x	1.43	2.15	2.86

In Figure 1, we have plotted the lines for males, females, and the combined group.

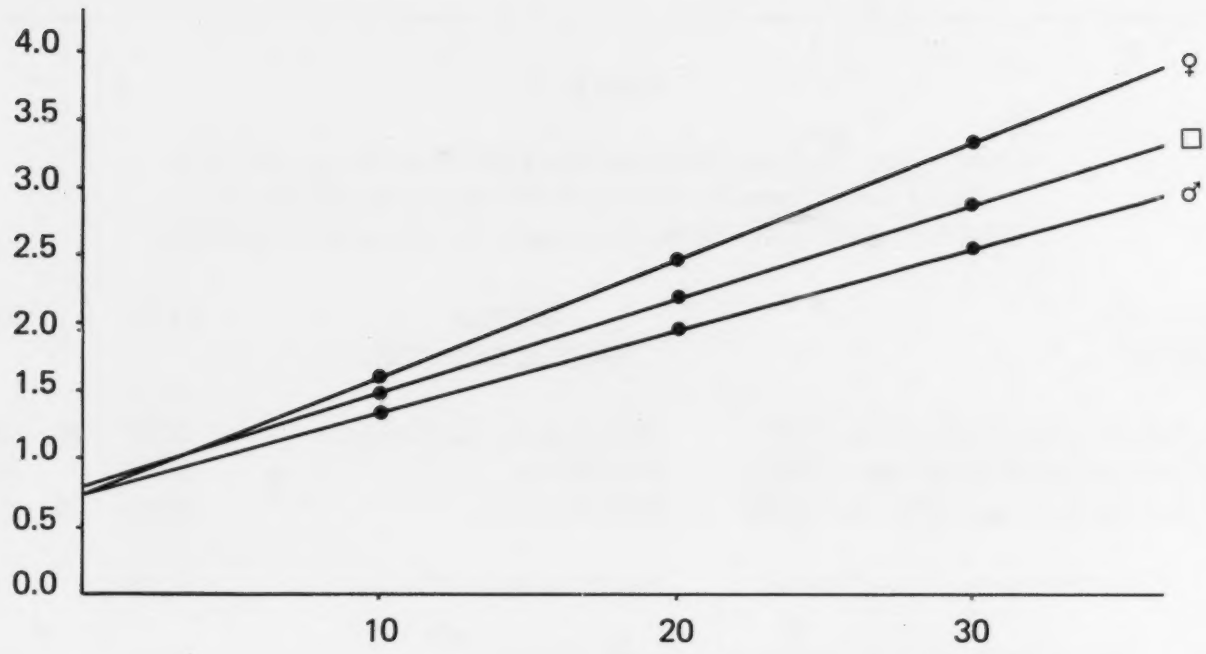


Fig. 1. Graph of Bayesian prediction lines for male, female, and overall groups for College 17.

The results presented here are typical. At all meaningful levels of ACT scores, the predicted GPA is higher in the female subpopulation and lower in the male subpopulation. The difference between these two values is large for the higher levels of ACT scores. The prediction line, derived from the total group, falls between that for the two subpopulations. Thus, it is clear that predictions based on the prediction function derived from the total group will, on the average, be too high for males and too low for females. One might well say that total group predictions discriminate against females, though other interpretations are possible.

Technical Details

The Bayesian model for regression in *m* groups as developed by Lindley (1970) is as follows:

a. Within each group we consider the linear regression

$$\mathcal{E}(y_{ij}) = \alpha_i + \sum_{h=1}^{\ell} \beta_{hi} x_{hij}$$

for i = 1, 2, ..., *m* colleges, j = 1, 2, ..., n_i persons within each college and h = 1, 2, ..., ℓ variables.

b. We write:

1. $\underset{\sim}{y}_i' = (y_{i1}, ..., y_{ij}, ..., y_{in_i})$ the vector of n_i observations of the criterion in College i .

2. $\underset{\sim}{\beta}_i' = (\alpha_i, \beta_{1i}, \beta_{2i}, ..., \beta_{hi}, ..., \beta_{\ell i})$ the vector of $\ell + 1$ regression weights for College i .

3. $\underset{\sim}{\beta}_{*}' = (\alpha_{\cdot}, \beta_{1\cdot}, \ldots, \beta_{\ell\cdot})$ where $\alpha_{\cdot} = m^{-1} \sum_{i=1}^{m} \alpha_i$ and $\beta_{h\cdot} = m^{-1} \sum_{i=1}^{m} \beta_{hi}$ the vector of average values, over colleges, of the α_i and the β_{hi}.

4.
$$\underset{\sim}{X}_i = \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_{1i1} & x_{1i2} & \cdots & x_{1ij} & \cdots & x_{1in_i} \\ \vdots & & & & & \\ x_{hi1} & x_{hi2} & \cdots & x_{hij} & \cdots & x_{hin_i} \\ \vdots & & & & & \\ x_{\ell i1} & x_{\ell i2} & \cdots & x_{\ell ij} & \cdots & x_{\ell in_i} \end{bmatrix},$$

the matrix of predictor scores for the n_i persons in the i-th college. The constant 1 has been put at the head of each column so that the α_i can be treated as regression weights β_{oi}.

5.
$$\underset{\sim}{\beta} = \begin{bmatrix} \underset{\sim}{\beta}_1' \\ \underset{\sim}{\beta}_2' \\ \vdots \\ \underset{\sim}{\beta}_i' \\ \vdots \\ \underset{\sim}{\beta}_m' \end{bmatrix} = \begin{bmatrix} \alpha_1 & \beta_{11} & \beta_{21} & \cdots & \beta_{h1} & \cdots & \beta_{\ell 1} \\ \alpha_2 & \beta_{12} & \beta_{22} & \cdots & \beta_{h2} & \cdots & \beta_{\ell 2} \\ \vdots & \vdots & \vdots & & & & \\ \alpha_i & \beta_{1i} & \beta_{2i} & \cdots & \beta_{hi} & \cdots & \beta_{\ell i} \\ \vdots & \vdots & \vdots & & & & \\ \alpha_m & \beta_{1m} & \beta_{2m} & \cdots & \beta_{hm} & \cdots & \beta_{\ell m} \end{bmatrix},$$

and

6.

$$\underset{\sim}{\beta}_* = \begin{bmatrix} \alpha. & \beta_{1\cdot} & \cdots & \beta_{h\cdot} & \cdots & \beta_{\ell\cdot} \\ \alpha. & \beta_{1\cdot} & \cdots & \beta_{h\cdot} & \cdots & \beta_{\ell\cdot} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \alpha. & \beta_{1\cdot} & \cdots & \beta_{h\cdot} & \cdots & \beta_{\ell\cdot} \end{bmatrix}$$

7. $\theta = m(\Sigma\phi_i^{-1})^{-1}$, the harmonic mean of the ϕ_i, where ϕ_i is the residual variance in the *i*-th college.

8. $\eta = \left(\prod_{i=1}^{m} \phi_i\right)^{1/m}$, the geometric mean of the ϕ_i.

9. κ = a small *positive* constant to be specified.

It is then assumed that these colleges and others that might be included in this analysis are exchangeable with respect to prior information, and, hence, by virtue of the de Finetti-Hewitt-Savage theorem (Lindley, in press), the available colleges can be treated as a random sample from some population of colleges. It is assumed that the $\underset{\sim}{\beta}_i'$ form a random sample from an $(\ell + 1)$-dimensional multivariate normal distribution with mean vector $\underset{\sim}{\mu}' = (\mu_0, \mu_1, \ldots, \mu_\ell)$, and with a dispersion matrix whose *inverse* has elements γ_{hk} (h, k = 0, 1, ..., ℓ); and we further assume the ϕ_i form a random sample such that $\nu\sigma^2/\phi_i$ is χ^2 on ν degrees of freedom. Further, the μ_i are taken to be independently uniform, a priori. The parameters ν and σ^2 are in effect taken to be independently and approximately, uniformly and log-uniformly distributed, though an additive constant, κ, is introduced to bound the posterior density in the region of the point at which all ϕ_i are equal. Effectively this stipulates that we have a priori reasons for knowing that all of the residual variances are not equal—hardly a controversial assumption.

The parameters γ_{hk} are assumed to have a Wishart distribution with parameters ν' and σ_{hk}. The matrix Σ is defined with typical element σ_{hk} and the elements of Σ^{-1} are denoted by σ^{hk}. Then $\mathcal{E}(\gamma_{hk}) = \sigma^{hk}$ so that σ_{hk} can be thought of as a prior estimate of the covariance across colleges of β_{hi} and β_{ki}. We then assume $\sigma_{hk} = 0$ and take $\sigma_{hh} = \sigma_h^2 > 0$. The constant ν' is a degrees of freedom parameter and effectively expresses the amount of prior information. When it is desirable to assume little prior information, it is convenient to take $\nu' = 1$ since the posterior will converge for $\nu' > 0$.

Proceeding to do the usual Bayesian integrations, which in this case required careful use of approximations to certain integrands, Lindley obtained the posterior joint density of $\underset{\sim}{B}$ and the ϕ_i. It did not seem possible to obtain expected values as point estimates so the modal estimate of the elements of $\underset{\sim}{B}$ and of the ϕ_i was sought by differentiation. What was sought was the matrix estimate $\tilde{\underset{\sim}{B}}$ of $\underset{\sim}{B}$ where $\tilde{\underset{\sim}{B}}$ has row vectors $\tilde{\underset{\sim}{\beta}}_i'$, containing the estimates for the *i*-th college and the estimates $\tilde{\phi}_i$ of ϕ_i, all from the joint posterior density.

Lindley's (1970) original solution involved the evaluation of all cofactors of large matrices. However, using the result that the matrix of cofactors is equal to the transpose of the product of the inverse of the original

matrix and the determinant of the matrix, the $2m$ Lindley equations can be stated in the following convenient forms:

$$\phi_i^{-1}(\underset{\sim}{y}_i'\underset{\sim}{X}_i') - \phi_i^{-1}\underset{\sim}{\beta}_i'(\underset{\sim}{X}_i\underset{\sim}{X}_i')$$

$$- (\nu' + m - 1)(\underset{\sim}{\beta}_i' - \underset{\sim}{\beta}_*')[\nu'\Sigma + (\underset{\sim}{B} - \underset{\sim}{B}_*)'(\underset{\sim}{B} - \underset{\sim}{B}_*)]^{-1} = \underset{\sim}{0}' \qquad (4.1)$$

$$- (n_i + 2) + \phi_i^{-1}(\underset{\sim}{y}_i'\underset{\sim}{y}_i) - 2\phi_i^{-1}\underset{\sim}{\beta}_i'(\underset{\sim}{X}_i\underset{\sim}{y}_i) + \phi_i^{-1}\underset{\sim}{\beta}_i'(\underset{\sim}{X}_i\underset{\sim}{X}_i')\underset{\sim}{\beta}_i$$

$$- \left(\frac{m+1}{m}\right)\left[1 - \frac{1}{\phi_i(\Theta^{-1} + \kappa)}\right]\left\{\frac{1}{\log[\eta(\Theta^{-1} + \kappa)]}\right\} = 0 \quad , \qquad (4.2)$$

for i = 2, ..., m .

Each of the first m equations is linear in the $\underset{\sim}{\beta}_i'$, and each of the second m equations is linear in ϕ_i^{-1} . Unfortunately all of the vectors $\underset{\sim}{\beta}_i'$ are involved in $\underset{\sim}{B}$ and $\underset{\sim}{B}^*$, which appear in each of the first m equations and all of the ϕ_i are involved in η and Θ, which appear in each of the second set of m equations. Therefore, an iterative solution is required. However, before proceeding with this, the validity of one of Lindley's assumptions is worth discussing.

The assumption that our prior information about the various regression coefficients is independent must be considered carefully. Generally, an independent prior assumption is standard in Bayesian work, and this is justified by asking oneself, "Suppose I know the value of parameter a , would this cause me to revise my prior distribution for parameter b ?" The answer is usually negative, justifying the use of independent priors for the set of regression coefficients $\beta_1, \beta_2, ..., \beta_\xi$. For simplicity, now consider a single predictor. In this case, α_i and β_i , the ordinate at the origin of x and the slope, respectively, of the *same* regression line, one cannot so lightly return a negative answer. For if α_i and β_i are independent, and α_i^* is the ordinate at some other value x^* of x , we have $\alpha_i^* = \alpha_i + \beta_i x^*$ and so the covariance

$$\sigma(\alpha_i^*, \beta_i) = x^*\sigma^2(\beta_i) \qquad (4.3)$$

showing that, unless β_i has zero variance, α_i^* and β_i are correlated and a fortiori not independent. Conversely, if α_i and β_i are correlated, we have

$$\sigma(\alpha_i^*, \beta_i) = \sigma(\alpha_i, \beta_i) + x^*\sigma^2(\beta_i) \qquad (4.4)$$

showing that the ordinate α_i^* at the point

$$x^* = \frac{-\sigma(\alpha_i, \beta_i)}{\sigma^2(\beta_i)} \tag{4.5}$$

is uncorrelated with β_i .

We are thus compelled, in formulating our prior distributions, to decide at what value x^* of x it is reasonable to assume that the ordinate and slope are independently distributed a priori. The Lindley equations are derived on the assumption that this point has been identified and taken as the origin of x . This difficulty is not an artifact of Bayesian methodology; it is intrinsic to the problem. For we wish, basically, to formalize our belief that the regression equations in the m groups will be similar. Thus, we expect the regression lines to be reasonably parallel (similar β's) and reasonably close together (similar α's). The latter statement is meaningless, however, except in the case when all the lines are *exactly* parallel (the case $\sigma^2(\beta_i) = 0$ noted above), unless we say *where* the lines are to be close together; any two nonparallel lines are close together—indeed meet—*somewhere*!

An obvious solution to this problem is to use (4.5) to scale x so that $\sigma(\alpha_i, \beta_i) = 0$ in the sample. In the multipredictor case, the extension of this method is to move the origin of x_n to the point which is the negative of the corresponding estimated coefficient in the regression of α on $\beta_1, \beta_2, \ldots, \beta_\ell$, across groups. A rescaling is then required after the analysis is completed. This was done for analyses reported here and we shall not discuss this issue further. A more complete discussion is given by Jackson, Novick, and Thayer (1970), where empirical work suggests that the precise determination of the x-value at which α and β are uncorrelated may not be important. The question of the prior covariance of the β_{hi}, β_{ki} is a far less sensitive issue. Some methods for adjusting for any available information are discussed by Jackson, Novick, and Thayer (1970), but considering the minimal weight being put on the prior distribution, this is unlikely to be of very great importance, unless the number of groups in the study is very small. It should, moreover, be stressed that the discussion in that paper refers to an earlier model in which the covariance of α and β was not a parameter of the model. In the current model, it is an explicit parameter and is, therefore, "estimated" by the Bayesian analysis. All we have to specify is that the prior expected value of the covariance is zero. The rescaling is thus a safety precaution, to ensure that this specification is reasonable: in many situations it may have no effect on the final estimates.

An interesting feature of the Lindley equations is that they apply in a limiting sense for some n_i approaching zero. For such colleges the solution is to take $\tilde{\underline{\beta}}_i' = \tilde{\underline{\beta}}_*'$, the vector of average values over colleges and $\tilde{\phi}_i = \eta$ the geometric mean of the Bayesian estimates of the ϕ_i . Since the Lindley equations were derived assuming noninformative priors, it follows that the procedure provides prediction weights despite the lack of either prior or sample information. The reason for this is that the collateral observations on the remaining colleges, in effect, provide an informative prior distribution for the colleges for which no observations are available.

At the other extreme, if any n_i is infinite, the Bayesian prediction weights correspond with the least squares weights. In all less extreme situations, each Bayesian weight is roughly a weighted average of the least squares weight and the average of the least squares weights across colleges; however, some adjustment does occur to balance off the several weights, including the intercept. In fact, the Bayesian result is a generalization of a formula due to Kelley (1927) for estimating true score (an expected value for a single experimental unit) as a

weighted average of the observed score for that unit and the population mean of the observed scores. This relationship is discussed in detail by Novick, Jackson, and Thayer (in press). Some simple classical analogs of the Bayesian solution are given by Jackson (in preparation). It is worth noting here that the special case $\ell = 0$ (no predictors) is precisely the Bayesian model II ANOVA solution to the Behrens-Fisher problem provided by Lindley (in press).

Solution of the Lindley equations is accomplished through an iterative procedure. The initial step in the data processing involves computing the usual within-group least squares regressions. The correlation matrix of the regression weights (including α) across colleges is then computed and the predictor variables are rescaled as indicated in the previous section. Since it is sometimes desirable to assume minimal prior information, we take $\nu' = 1$. When m is small this may be unsatisfactory. The off-diagonal elements of Σ are taken to be zero indicating that our prior beliefs about the covariances of the regression coefficients are independent. The diagonal elements are generally taken to correspond to prior knowledge or as was done here can be "bootstrapped" from the data using sample estimates. Such values have very little effect on the final result. It is only necessary to take our prior estimates not too near zero.

The least squares estimates are used as starting points for the iterations. In the first set of Lindley equations (4.1), these values are substituted for ϕ_i, $\underset{\sim}{\beta}_*$, $\underset{\sim}{B}$, and $\underset{\sim}{B}_*$ (for all occurrences of $\underset{\sim}{\beta}_i$ and ϕ_i except where $\underset{\sim}{\beta}_i$ appears explicitly). Then we have m sets of $(\ell + 1)$ linear equations in $(\ell + 1)$ unknowns, one for each of the m vectors $\underset{\sim}{\beta}_i$ in (4.1). In the second set of equations (4.2), the least squares estimates for ϕ_i are used to compute η and Θ. Then we have m equations, each linear in ϕ_i^{-1}. Each of the first sets is solved and the obtained values put immediately back into the set twice to stabilize the solution. The resulting set of $m(\ell + 1)$ estimates is then put into the equations for the ϕ_i^{-1} and these are solved and iterated a further two times. This constitutes one cycle. The resulting set of $m(\ell + 2)$ estimates can then be put back into a second cycling. Estimates reported in the Results section of this report were based on 200 such cycles. A listing and detailed description of the program is available from any of the authors.

References

Gleazer, E. J., Jr. (Ed.) *American junior colleges.* Washington, D.C.: American Council on Education, 1967.

Jackson, P. H. The estimation of many parameters—some simple approximations. ACT *Technical Bulletin No. 2.* Iowa City, Iowa: The American College Testing Program, 1971.

Jackson, P. H., Novick, M. R., & Thayer, D. T. *Bayesian inference and the classical test theory model, II. Validity and prediction.* Research Bulletin 70-32. Princeton, N.J.: Educational Testing Service, 1970.

Kelley, T. L. *The interpretation of educational measurements.* Yonkers on Hudson, N.Y.: World Book, 1927.

Lindley, D. V. *A Bayesian solution for some educational prediction problems, III.* Research Bulletin 70-33. Princeton, N.J.: Educational Testing Service, 1970.

Lindley, D. V. The estimation of many parameters. *Proceedings of the Waterloo Conference on the Foundations of Statistics,* in press.

Novick, M. R., Jackson, P. H., & Thayer, D. T. Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika,* in press.

ACT Research Reports

This report is Number 42 in a series published by the Research and Development Division of The American College Testing Program. The first 26 research reports have been deposited with the American Documentation Institute, ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Photocopies and 35 mm. microfilms are available at cost from ADI; order by ADI Document number. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress. Beginning with Research Report No. 27, the reports have been deposited with the National Auxiliary Publications Service of the American Society for Information Science (NAPS), c/o CCM Information Sciences, Inc., 22 West 34th Street, New York, New York 10001. Photocopies and 35 mm. microfilms are available at cost from NAPS. Order by NAPS Document number. Advance payment is required. Printed copies may be obtained, if available, from the Research and Development Division, The American College Testing Program.

The reports since January 1969 in this series are listed below. A listing of previous reports is included in each of several items published by The American College Testing Program: *Your College Freshmen* (pp. 158-160), *Your College-Bound Students* (pp. 107-109). A complete list of the reports can be obtained by writing to the Research and Development Division, The American College Testing Program, P. O. Box 168, Iowa City, Iowa 52240.

No. 28 *A Description of Graduates of Two-Year Colleges,* by L. L. Baird, J. M. Richards, Jr., & L. R. Shevel (NAPS No. 11306; photo, $3.00; microfilm, $1.00)

No. 29 *An Empirical Occupational Classification Derived from a Theory of Personality and Intended for Practice and Research,* by J. L. Holland, D. R. Whitney, N. S. Cole, & J. M. Richards, Jr. (NAPS No. 00505; photo, $3.00; microfilm, $1.00)

No. 30 *Differential Validity in the ACT Tests,* by N. S. Cole (NAPS No. 00722; photo, $3.00; microfilm, $1.00)

No. 31 *Who Is Talented? An Analysis of Achievement,* by C. F. Elton, & L. R. Shevel (NAPS No. 00723; photo, $3.00; microfilm, $1.00)

No. 32 *Patterns of Educational Aspiration,* by L. L. Baird (NAPS No. 00920; photo, $3.00; microfilm, $1.00)

No. 33 *Can Financial Need Analysis Be Simplified?* by M. D. Orwig, & P. K. Jones (NAPS No. 01210; photo, $5.00; microfilm, $3.00)

No. 34 *Research Strategies in Studying College Impact,* by K. A. Feldman (NAPS No. 01211; photo, $5.00; microfilm, $2.00)

No. 35 *An Analysis of Spatial Configuration and Its Application to Research in Higher Education,* by N. S. Cole, & J. W. L. Cole (NAPS No. 01212; photo, $5.00; microfilm, $2.00)

No. 36 *Influence of Financial Need on the Vocational Development of College Students,* by A. R. Vander Well (NAPS No. not available at this time.)

No. 37 *Practices and Outcomes of Vocational-Technical Education in Technical and Community Colleges,* by T. G. Gartland, & J. F. Carmody (NAPS No. not available at this time.)

No. 38 *Bayesian Considerations in Educational Information Systems,* by M. R. Novick (NAPS No. not available at this time.)

No. 39 *Interactive Effects of Achievement Orientation and Teaching Style on Academic Achievement,* by G. Domino (NAPS No. not available at this time.)

No. 40 *An Analysis of the Structure of Vocational Interests,* by N. S. Cole, & G. R. Hanson (NAPS No. not available at this time.)

No. 41 *How Do Community College Transfer and Occupational Students Differ?* by E. J. Brue, H. B. Engen, & E. J. Maxey (NAPS No. not available at this time.)